

Penalized integrative analysis under the accelerated failure time model

Qingzhao Zhang¹, Sanguo Zhang¹, Jin Liu², Jian Huang³ and Shuangge Ma^{4*}

¹School of Mathematical Sciences, University of Chinese Academy of Sciences

²Center of Quantitative Medicine, Duke-NUS School of Medicine

³Department of Statistics and Actuarial Science, University of Iowa

⁴Department of Biostatistics, Yale University

**email: shuangge.ma@yale.edu*

January 13, 2015

Running Title: Integrative analysis under AFT model

Abstract

For survival data with high-dimensional covariates, results generated in the analysis of a single dataset are often unsatisfactory because of the small sample size. Integrative analysis pools raw data from multiple independent studies with comparable designs, effectively increases sample size, and has better performance than meta-analysis and single-dataset analysis. In this study, we conduct integrative analysis of survival data under the accelerated failure time (AFT) model. The sparsity structures of multiple datasets are described using the homogeneity and heterogeneity models. For variable selection under the homogeneity model, we adopt group penalization approaches. For variable selection under the heterogeneity model, we use composite penalization and sparse group penalization approaches. As a major advancement from the existing studies, the asymptotic selection and estimation properties are rigorously established. Simulation study is conducted to compare different penalization methods and against alternatives. We also analyze four lung cancer prognosis datasets with gene expression measurements.

Keywords: Integrative analysis; Homogeneity and heterogeneity models; Penalized selection; Consistency properties.

1 Introduction

In survival studies, data with high-dimensional covariates are now commonly encountered. A lung cancer prognosis study with gene expression measurements is presented in this article,

and more are available in the literature. With such “large p , small n ” data, results generated in the analysis of a single dataset are often unsatisfactory because of the small sample size (Guerra and Goldstein, 2009; Liu et al., 2013; Ma et al., 2011b). For outcomes of common interest, there are often multiple independent studies with comparable designs. This makes it possible to pool multiple datasets, increase sample size, and improve over single-dataset analysis. As a family of multi-dataset analysis methods, integrative analysis methods pool and analyze raw data from multiple studies and outperform classic meta-analysis methods, which analyze multiple datasets separately and then combine summary statistics.

In this article, we conduct the integrative analysis of multiple independent survival datasets under the accelerated failure time (AFT) model. The analysis goal is to identify, out of a large number of measured covariates, important markers associated with survival. For such a purpose, we adopt penalization, which has been the choice of many high-dimensional studies. A large number of penalization methods have been developed for single-dataset analysis. However because of the multi-dataset settings and heterogeneity across datasets, they are not applicable to integrative analysis. The sparsity structures of multiple datasets can be described using the homogeneity and heterogeneity models. Different models demand marker selection with different properties and hence different methods. This makes integrative analysis even more complicated. Penalization methods for integrative analysis have been developed (Liu et al., 2013; Ma et al., 2011b), however, in an unsystematic manner.

This study advances from the existing ones in the following aspects. First, it advances from single-dataset analysis and meta-analysis by conducting integrative analysis of multiple heterogeneous datasets. Second, it conducts more systematic investigation than the existing integrative analysis studies such as Liu et al. (2013); Ma et al. (2011b). *More importantly, it rigorously establishes the selection and estimation properties which have not been previously examined.* The theoretical development is nontrivial because of data complexity, model settings, and penalties. Third, the properties of composite penalization and sparse group penalization have not been studied for single-dataset analysis under the AFT model. Thus our study can also provide insights for single-dataset penalization methods. Fourth, this study also advances from the existing studies by conducting systematic simulations and direct comparisons of multiple methods.

Data and model settings are described in Section 2. Penalized integrative analyses under the homogeneity and heterogeneity models are investigated in Section 3 and 4 respectively. We conduct numerical study in Section 5. The article concludes with discussions in Section 6. Technical details and additional analysis results are provided in Appendix.

2 Integrative analysis under AFT model

Consider the integrative analysis of survival data from M independent studies. In study $m (= 1, \dots, M)$ with n_m iid subjects, let $\mathbf{T}^m = (T_1^m, \dots, T_{n_m}^m)^\top$ be the logarithm of failure times and $\mathbf{X}^m \in R^{n_m \times p_m}$ be the predictor matrix. Assume the AFT model

$$\mathbf{T}^m = \mathbf{X}^m \boldsymbol{\beta}^m + \boldsymbol{\epsilon}^m. \quad (1)$$

β^m is the vector of regression coefficients, and ϵ^m is the vector of random errors. With proper normalization, the intercept term has been omitted. Assume that all datasets measure the same set of covariates. Then $p_1 = \dots = p_M = p$. When different datasets have mismatched covariate sets, a rescaling approach (Ma et al., 2011a; Liu et al., 2013) can be adopted. The proposed approaches are then applicable with minor modifications.

Let $\beta = (\beta^1, \dots, \beta^M) = (\beta_1, \dots, \beta_p)^\top$, where $\beta_j = (\beta_j^1, \dots, \beta_j^M)^\top$ consists of the coefficients of variable j in all M datasets. Moreover, write $\beta = (\beta_{ij})_{p \times M}$ with its true value β^* , where $\beta_{ij} = \beta_i^j$. With the heterogeneity across datasets, β_j^m is not necessarily equal to β_j^k for $m \neq k$. Under right censoring, one observes (Y^m, δ^m, X^m) with $Y^m = T^m \wedge C^m$, where C^m is the vector of log censoring times, and $\delta^m = 1\{T^m \leq C^m\}$.

When the distribution of random errors is unknown, there are multiple estimation approaches (Ying, 1993). We adopt the weighted least squares (LS) approach (Stute, 1993), which has the lowest computational cost and is desirable with high-dimensional data. Let \hat{F}^m be the Kaplan-Meier estimator of the distribution function F^m of T^m . Let $Y_{(1)}^m \leq \dots \leq Y_{(n_m)}^m$ be the order statistics of Y_i^m 's. \hat{F}^m can be written as $\hat{F}^m(y) = \sum_{i=1}^{n_m} \omega_i^m 1\{Y_i^m \leq y\}$, where ω_i^m 's are expressed as $\omega_1^m = \frac{\delta_{(1)}^m}{n_m}$ and $\omega_i^m = \frac{\delta_{(i)}^m}{n_m - i + 1} \prod_{j=1}^{i-1} \left(\frac{n_m - j}{n_m - j + 1} \right)^{\delta_{(j)}^m}$, $i = 2, \dots, n_m$. Here $\delta_{(1)}^m, \dots, \delta_{(n_m)}^m$ are the associated censoring indicators of the ordered Y_i^m 's. Denote $W_m = \text{diag}\{n_m \omega_1^m, \dots, n_m \omega_{n_m}^m\}$. Then for the M datasets combined, the weighted LS approach is to minimize

$$\tilde{L}(\beta) = \frac{1}{2n} \sum_{m=1}^M (Y^m - X^m \beta^m)^\top W_m (Y^m - X^m \beta^m). \quad (2)$$

Note that the components of Y^m and X^m need to be sorted. Assume that:

- [CONDITION 1] (a) The n_m components of ϵ^m are i.i.d. and sub-Gaussian with noise level σ_m . That is, for all vector ν with $\|\nu\|_2 = 1$ and any $t \geq 0$, $P(|\nu^\top \epsilon^m| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma_m^2}\right)$.
(b) ϵ^m is independent of W_m .

The total sample size is $n = \sum_{m=1}^M n_m$. The important predictor index sets of M datasets are respectively labeled as S_1, \dots, S_M . Then $S = \bigcup_{m=1}^M S_m$ denotes the important set with its corresponding variables important in at least one dataset. Let S^c and $|S|$ denote the complement and cardinality of set S , respectively. Let $\mathcal{A} = \{(i, j) : \beta_{ij}^* \neq 0\}$ and $\mathcal{B} = \{(i, j) : i \in S, j = 1, \dots, M\}$. Let $\beta_{\mathcal{A}}$ and $\beta_{\mathcal{B}}$ denote the components of β indexed by \mathcal{A} and \mathcal{B} , respectively. For a $p \times 1$ vector v and index set $I \subset \{1, \dots, p\}$, let v_I denote the components of v indexed by I . Moreover, let $X^{m,i}$ denotes the transposition of the i th row of X^m . Then for any index set $I \subset \{1, \dots, p\}$, $X_I^m = (X_I^{m,1}, \dots, X_I^{m,n_m})^\top$.

2.1 Homogeneity and heterogeneity models

The sparsity structure of β can be described using the homogeneity and heterogeneity models. Under the homogeneity model, β^m 's have the same sparsity structure. That is,

$I(\beta_j^m = 0) = I(\beta_j^k = 0)$ for all (m, k, j) 's. The intuition is that if the M datasets are “close enough”, then the same set of markers should be identified in all datasets. Under this model, we only need to determine whether a covariate is important or not, that is, only one level of selection is needed. With the (sometimes great) differences across datasets, the homogeneity model may be too restricted. As an alternative, the heterogeneity model allows different datasets to have different sparsity structures. It includes the homogeneity model as a special case and can be more flexible. Under this model, we need to determine whether a covariate is associated with any response at all. In addition, for an important covariate, we need to determine in which datasets it is important. That is, a two-level selection is needed.

3 Integrative analysis under the homogeneity model

Under this model, one-level selection is needed and can be achieved using group penalization. In terms of formulation and computation, the development of group penalization methods in integrative analysis share some similarity with that in single-dataset analysis (Bühlmann and van de Geer, 2011). However, with the significantly different data settings and adoption of the AFT model, the theoretical development has significant differences.

3.1 Group LASSO

Consider the group LASSO penalized objective function

$$L(\beta) = \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}^m \beta^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}^m \beta^m) + \lambda \sum_{j=1}^p \|\beta_j\|_2, \quad (3)$$

where λ is the tuning parameter and $\|\beta_j\|_2 = [(\beta_j^1)^2 + \dots + (\beta_j^M)^2]^{1/2}$.

For set S , define the estimate $\hat{\beta}_S = (\hat{\beta}_S^1, \dots, \hat{\beta}_S^M)$ as

$$\hat{\beta}_S = \arg \min_{\beta_S} \left\{ \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}_S^m \beta_S^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}_S^m \beta_S^m) + \lambda \sum_{j \in S} \|\beta_j\|_2 \right\}. \quad (4)$$

For group LASSO to be able to consistently identify the true sparsity structure, there needs a local solution $\hat{\beta}^{glasso} = \{\hat{\beta}_B^{glasso}, \hat{\beta}_{B^c}^{glasso}\}$ for (3), where $\hat{\beta}_B^{glasso} = \hat{\beta}_B$ and $\hat{\beta}_{B^c}^{glasso} = 0$. Define

$$\begin{aligned} \bar{\rho}_2^m &= \lambda_{\max}\{n_m^{-1} \mathbf{X}_S^{m\top} W_m^2 \mathbf{X}_S^m\}, \quad \underline{\rho}_1^m = \lambda_{\min}\{n_m^{-1} \mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m\} \\ \Lambda_m &= \max_j \{n_m^{-1} \mathbf{X}_j^{m\top} W_m^2 \mathbf{X}_j^m\}, \quad \psi_m = \|\mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1}\|_\infty. \end{aligned}$$

Theorem 1 *Consider the estimator defined by minimizing (3). Under Condition 1,*

1. There exists a local minimizer $\hat{\beta}_{\mathcal{B}}$ of (4) such that

$$\Pr \left\{ \|\hat{\beta}_S^m - \beta_S^{m*}\|_2 \leq \lambda \sqrt{|S|} \frac{4}{\rho_1^m} \frac{n}{n_m}, m = 1, \dots, M \right\} \geq 1 - \sum_{m=1}^M \exp \left(-\frac{\lambda^2 |S| n^2}{2\sigma_m^2 \bar{\rho}_2^m n_m} \right).$$

2. Assume the ir-representable conditions $\psi_m \leq D_m < 1$. $\hat{\beta}^{glasso} = \{\hat{\beta}_{\mathcal{B}}^{glasso}, \hat{\beta}_{\mathcal{B}^c}^{glasso}\}$ with $\hat{\beta}_{\mathcal{B}}^{glasso} = \hat{\beta}_{\mathcal{B}}, \hat{\beta}_{\mathcal{B}^c}^{glasso} = 0$ is a local minimizer of (3) with probability at least

$$1 - \sum_{m=1}^M \exp \left(-\frac{\lambda^2 |S| n^2}{2\sigma_m^2 \bar{\rho}_2^m n_m} \right) - 2p \sum_{m=1}^M \exp \left\{ -\frac{n^2 \lambda^2 (1 - D_m)^2}{2n_m \Lambda_m \sigma_m^2 (1 + D_m)^2} \right\}.$$

In single-dataset analysis, Zhao and Yu (2006) and followup studies establish selection consistency under the ir-representable condition. Under a similar condition for individual datasets, integrative analysis also has selection consistency.

With the probability bounds in Theorem 1, we can obtain a more straightforward understanding of the penalized estimators and derive the following result.

Corollary 1 Suppose that for $m = 1, \dots, M$, $\rho_1^m, \bar{\rho}_2^m$, and Λ_m are bounded away from zero and infinity. Assume that $n/n_m = O(1)$, $|S| \ll n$, and $\log p = O(n^\alpha)$ with $\alpha < 1$. Under Condition 1 and the ir-representable conditions in Theorem 1, if $|S|^{-1/2} \min_{j \in S} \|\beta_j^*\|_2 \gg \lambda \gg n^{\frac{\alpha-1}{2}}$, then group LASSO can identify the true sparsity structure and $\|\hat{\beta}_S^m - \beta_S^{m*}\|_2 = O_p(\lambda \sqrt{|S|})$, $m = 1, \dots, M$.

Remark 1 It is known that in single-dataset analysis the group LASSO is group selection consistent under some variants of the ir-representable condition. See Huang et al. (2012) and others for reference. Similar conditions are needed in the integrative analysis with group LASSO. The conditions in Corollary 1 on $\rho_1^m, \bar{\rho}_2^m$, and Λ_m are on the design matrixes and censoring probabilities. Corollary 1 shows that even when the group LASSO can identify the true sparsity structure, λ should be much large than $n^{-1/2}$, leading to $\|\hat{\beta}_S^m - \beta_S^{m*}\|_2 \gg \sqrt{|S|/n}$.

3.2 Concave 2-norm group selection

Consider penalization built on concave penalties. Notable examples of concave penalty include SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). For $t \geq 0$, the SCAD penalty has first order derivative $p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}$, for some $a > 2$. The MCP has derivative $p'_\lambda(t) = \lambda \left(1 - \frac{t}{a\lambda} \right)_+$, for some $a > 1$. Consider the objective function

$$L(\beta) = \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}^m \beta^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}^m \beta^m) + \sum_{j=1}^p p_\lambda(\|\beta_j\|_2), \quad (5)$$

where the penalty $p_\lambda(\cdot)$ satisfies:

[CONDITION 2] $\lambda^{-1}p_\lambda(t)$ is concave in $t \in [0, \infty)$ with a continuous derivative $\lambda^{-1}p'_\lambda(t)$ satisfying $\lambda^{-1}p'_\lambda(0+) \in (0, \infty)$. In addition, $\lambda^{-1}p'_\lambda(t)$ is increasing in $\lambda \in (0, +\infty)$, and $\lambda^{-1}p'_\lambda(0+)$ is independent of λ .

[CONDITION 3] $\theta = \inf \left\{ \frac{t}{\lambda} : \lambda^{-1}p'_\lambda(t) = 0, t \geq 0 \right\}$ is bounded.

Remark 2 Condition 2 is also considered by Fan and Lv (2011). LASSO, SCAD, and MCP all satisfy this condition. Condition 3 is added to guarantee unbiasedness. LASSO does not satisfy Condition 3 since $\lambda^{-1}p'_\lambda(t) = 1$ leads to $\theta = \infty$, while SCAD and MCP satisfy with $\theta = a$. Another approach that has been studied is the 2-norm group bridge (Ma et al., 2012). Under certain conditions, its selection consistency is established in Ma et al. (2011a). Note that the bridge penalty does not satisfy Condition 3 and needs to be separately investigated.

Consider the properties of concave 2-norm group penalization. Define the oracle estimator as $\hat{\beta}^{oracle} = \{\hat{\beta}_B^{oracle}, \hat{\beta}_{B^c}^{oracle}\}$ with $\hat{\beta}_B^{oracle} = \tilde{\beta}_B$ and $\hat{\beta}_{B^c}^{oracle} = 0$, where

$$\tilde{\beta}_B = \arg \min_{\beta_S} \left\{ \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}_S^m \beta_S^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}_S^m \beta_S^m) \right\}. \quad (6)$$

Theorem 2 Under Condition 1-3, consider the estimator defined by minimizing (5).

1. For any $R_m < \sqrt{\frac{n_m}{|S|}}$, we have

$$\Pr \left(\|\tilde{\beta}_S^m - \beta_S^{m*}\| \leq \sqrt{\frac{|S|}{n_m}} R_m, m = 1, \dots, M \right) \geq 1 - \sum_{m=1}^M \exp \left\{ -R_m^2 \frac{|S|(\rho_1^m)^2}{8\bar{\rho}_2^m \sigma_m^2} \right\}.$$

2. Suppose $\lambda < \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\theta}$ and $R_m^\dagger \leq \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\sqrt{M}} \sqrt{\frac{n_m}{|S|}}$. Then with probability at least

$$1 - \sum_{m=1}^M \exp \left\{ -\frac{|S|(\rho_1^m)^2}{8\bar{\rho}_2^m \sigma_m^2} R_m^{\dagger 2} \right\} - 2p \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_\lambda'^2(0+)}{2n_m \Lambda_m \sigma_m^2 (1 + \psi_m)^2} \right\},$$

$\hat{\beta}^{oracle}$ is a local minimizer of (5).

Theorem 2 can be used to derive the following asymptotic result.

Corollary 2 Suppose that for $m = 1, \dots, M$, $\rho_1^m, \bar{\rho}_2^m$ and Λ_m are bounded away from zero and infinity, $n/n_m = O(1)$, $|S| \ll n$, $\log p = \tilde{O}(n^\alpha)$ with $\alpha < 1$, and $\psi_m = O(n^{\alpha_1})$ with $\alpha_1 \in [0, 1/2)$. Under Condition 1-3, if $\lambda < \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\theta}$ and $\lambda \gg n^{\frac{\alpha-1}{2} + \alpha_1}$, then the concave 2-norm group selection can identify the true sparsity structure and $\|\hat{\beta}_S^m - \beta_S^{m*}\|_2 = O_p(\sqrt{\frac{|S|}{n_m}})$.

Remark 3 When the concave penalty is used, the upper bound of ψ_m can grow to ∞ at rate $O(n^{\alpha_1})$. In contrast, the group LASSO needs the ir-representable conditions. Moreover, the group LASSO yields a larger bias than the concave 2-norm group selection.

4 Integrative analysis under the heterogeneity model

Under this model, two-level selection is needed and can be achieved using composite penalization and sparse group penalization. Properties of composite penalization have been studied in single-dataset analysis, however, under much simpler data and model settings. For sparse group penalization built on concave penalties, properties have not been established for single-dataset analysis.

Define the oracle estimator $\check{\beta} = \{\check{\beta}_A, 0\}$ where

$$\check{\beta}_A = \arg \min_{\beta_A} \left\{ \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}_{S_m}^m \beta_{S_m}^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}_{S_m}^m \beta_{S_m}^m) \right\}. \quad (7)$$

Define $\bar{\rho}_2^{*m} = \lambda_{\max}\{n_m^{-1} \mathbf{X}_{S_m}^{m\top} W_m^2 \mathbf{X}_{S_m}^m\}$, $\underline{\rho}_1^{*m} = \lambda_{\min}\{n_m^{-1} \mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m\}$ and $\psi_m^* = \|\mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m (\mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m)^{-1}\|_\infty$.

Theorem 3 Consider the estimator defined in (7). Under Condition 1-3, we have

$$\Pr \left\{ \|\check{\beta}_{S_m}^m - \beta_{S_m}^{m*}\|_2 \leq \sqrt{\frac{|S_m|}{n_m}} C_m, m = 1, \dots, M \right\} \geq 1 - \sum_{m=1}^M \exp \left\{ -C_m^2 \frac{|S_m| (\underline{\rho}_1^{*m})^2}{8 \bar{\rho}_2^{*m} \sigma_m^2} \right\}$$

with $C_m < \sqrt{\frac{n_m}{|S_m|}}$.

Corollary 3 Suppose that for $m = 1, \dots, M$, $\underline{\rho}_1^{*m}$ and $\bar{\rho}_2^{*m}$ are bounded away from zero and infinity, $n/n_m = O(1)$, and $|S| \ll n$. Under Condition 1-3, $\|\check{\beta}_{S_m}^m - \beta_{S_m}^{m*}\|_2 = O_p(\sqrt{\frac{|S_m|}{n_m}})$ for $m = 1, \dots, M$.

4.1 Composite penalization

Consider the objective function

$$L(\beta) = \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}^m \beta^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}^m \beta^m) + \sum_{j=1}^p p_{O, \lambda_O} \left(\sum_{m=1}^M p_{I, \lambda_I}(|\beta_j^m|) \right), \quad (8)$$

where the outer penalty $p_{O, \lambda_O}(\cdot)$ determines the overall importance of a variable, and the inner penalty $p_{I, \lambda_I}(\cdot)$ determines its individual importance. λ_O and λ_I are tuning parameters. A specific example is the composite MCP (cMCP) where both p_{O, λ_O} and p_{I, λ_I} are MCP.

[CONDITION 4] $\theta_O = \inf \left\{ \frac{|t|}{\lambda_O} : \frac{p'_{O,\lambda_O}(|t|)}{\lambda_O} = 0 \right\}$ and $\theta_I = \inf \left\{ \frac{|t|}{\lambda_I} : \frac{p'_{I,\lambda_I}(|t|)}{\lambda_I} = 0 \right\}$ are bounded.

Denote $J^{-m} = \max \left\{ \sum_{i \neq m}^M I(\beta_j^i \neq 0), j \in S - S_m \right\}$ and $f_I^{\max} = \max_t p_{I,\lambda_I}(t)$.

Theorem 4 Consider the minimizer of (8). Assume Condition 1-2 and 4. Set

$$C_m^\dagger \leq \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2} \sqrt{\frac{n_m}{|S_m|}}, \quad \lambda_I < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_I}, \quad \lambda_O \theta_O > f_I^{\max} \max_m (J^{-m}).$$

Then $\check{\beta}$ is a local minimizer with probability at least $1 - \tau_2$, where

$$\begin{aligned} \tau_2 = & \sum_{m=1}^M \exp \left\{ -C_m^{\dagger 2} \frac{|S_m|(\underline{\rho}_1^{*m})^2}{8\bar{\rho}_2^{*m}\sigma_m^2} \right\} + 2|S| \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{I,\lambda_I}^{\prime 2}(0+) p_{O,\lambda_O}^{\prime 2}(J^{-m} f_I^{\max})}{2n_m \bar{\rho}_2^{*m} \sigma_m^2 (1 + \psi_m^*)^2} \right\} \\ & + 2(p - |S|) \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{I,\lambda_I}^{\prime 2}(0+) p_{O,\lambda_O}^{\prime 2}(0+)}{2n_m \Lambda_m \sigma_m^2 (1 + \psi_m^*)^2} \right\}. \end{aligned}$$

This theorem establishes the consistency of composite penalized estimates. A simplified statement is provided in the following corollary.

Corollary 4 Suppose that for $m = 1, \dots, M$, $\underline{\rho}_1^{*m}, \bar{\rho}_2^{*m}$, and Λ_m are bounded away from zero and infinity, $n/n_m = O(1)$, $|S| \ll n$, $\log p = O(n^\alpha)$ with $\alpha < 1$, and $\psi_m^* = O(n^{\alpha_1})$ with $\alpha_1 \in [0, 1/2)$. Under Condition 1,2 and 4, if $\lambda_I < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_I}$, $\lambda_O \theta_O = M f_I^{\max}$, and $\lambda_I \lambda_O \gg n^{\frac{\alpha-1}{2} + \alpha_1}$, composite penalization can achieve the two-level selection consistency.

Remark 4 Liu et al. (2014) also suggests the composition of MCP and LASSO. We conjecture that it is estimation consistent, can consistently identify the overall importance of variables, but in general is not consistent at the individual level.

4.2 Sparse group penalization

Consider the objective function

$$L(\beta) = \frac{1}{2n} \sum_{m=1}^M (\mathbf{Y}^m - \mathbf{X}^m \beta^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}^m \beta^m) + \sum_{j=1}^p p_{1,\lambda_1}(\|\beta_j\|_2) + \sum_{j=1}^p \sum_{m=1}^M p_{2,\lambda_2}(|\beta_j^m|). \quad (9)$$

λ_1 and λ_2 are tuning parameters. Here the penalty is the sum of group and individual penalties. The first penalty determines the overall importance of a variable, and the second penalty determines its individual importance.

Consider penalties p_{1,λ_1} and p_{2,λ_2} that satisfy Condition 2 and 4 with bounded constants θ_1 and θ_2 . Consider the estimator defined by minimizing (9).

Theorem 5 Suppose that Condition 1-2 and 4 hold. Set

$$C_m^\dagger \leq \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2} \sqrt{\frac{n_m}{|S_m|}}, \quad \lambda_1 < \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\theta_1}, \quad \lambda_2 < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_2}.$$

Then $\check{\beta}$ is a local minimizer with probability at least $1 - \tau_3$, where

$$\begin{aligned} \tau_3 = & \sum_{m=1}^M \exp \left\{ -C_m^{\dagger 2} \frac{|S_m|(\rho_1^{*m})^2}{8\bar{\rho}_2^{*m}\sigma_m^2} \right\} + 2|S| \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{2,\lambda_2}'^2(0+)}{2n_m \bar{\rho}_2^{*m} \sigma_m^2 (1 + \psi_m^*)^2} \right\} \\ & + 2(p - |S|) \sum_{m=1}^M \exp \left\{ -\frac{n^2 [p_{1,\lambda_1}'(0+) + p_{2,\lambda_2}'(0+)]^2}{2n_m \Lambda_m \sigma_m^2 (1 + \psi_m^*)^2} \right\}. \end{aligned}$$

That is, the sparse group penalization also enjoys the consistency properties. For theoretical purpose, p_{1,λ_1} and p_{2,λ_2} do not need to take the same form. However using the same p_{1,λ_1} and p_{2,λ_2} may facilitate computation. We then derive the following asymptotic result.

Corollary 5 Suppose that for $m = 1, \dots, M$, $\rho_1^{*m}, \bar{\rho}_2^{*m}$, and Λ_m are bounded away from zero and infinity, $n/n_m = O(1)$, $|S| \ll n$, $\log p = O(n^\alpha)$ with $\alpha < 1$, and $\psi_m^* = O(n^{\alpha_1})$ with $\alpha_1 \in [0, 1/2)$. Under Condition 1-2 and 4, if $\lambda_1 < \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\theta_1}$, $\lambda_2 < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_2}$, $\lambda_1 \gg n^{-\frac{1}{2} + \alpha_1}$ and $\lambda_1 + \lambda_2 \gg n^{\frac{\alpha-1}{2} + \alpha_1}$, then the sparse group penalization achieves the two-level selection consistency.

5 Numerical study

5.1 Computation

With the weighted LS approach, the loss function (2) has a least squares form. In single-dataset analysis with a LS loss, multiple computational algorithms have been developed for group penalization, composite penalization, and sparse group penalization (Friedman et al., 2010; Breheny and Huang, 2009; Liu et al., 2014). Here we adopt the existing gradient descent algorithms with minor modifications. Convergence properties can be derived following Breheny and Huang (2011) and references therein. Details are omitted here. The penalization methods involve the tuning parameter $\lambda(\lambda_I, \lambda_O, \lambda_1, \lambda_2)$. The theorems provide results on the asymptotic order. MCP also involves the additional regularization parameter a . Following the literature, we consider a small number of values for a , in particular including 1.8, 3, 6 and 10. In numerical study, we use 5-fold cross validation for tuning parameter selection.

5.2 Simulation

We simulate three datasets, each with 100 subjects. For each subject, we simulate 1,000 covariates. The covariates have a joint normal distribution, with marginal means equal

to zero and variances equal to one. Consider two correlation structures. The first is the auto-regressive (AR) correlation, where covariates j and k have correlation coefficient $\rho^{|j-k|}$. $\rho = 0.2, 0.5$, and 0.8 , corresponding to weak, moderate, and strong correlations, respectively. The second is the banded correlation. Here three scenarios are considered. Under the first scenario, covariates j and k have correlation coefficient 0.3 if $|j - k| = 1$ and 0 otherwise. Under the second scenario, covariates j and k have correlation coefficient 0.6 if $|j - k| = 1$, 0.3 if $|j - k| = 2$, and 0 otherwise. Under the third scenario, covariates j and k have correlation coefficient 0.6 if $|j - k| = 1$, 0.3 if $|j - k| = 2$, 0.15 if $|j - k| = 3$, and 0 otherwise. Both the homogeneity and heterogeneity models are simulated. Under the homogeneity model, all three datasets share the same twenty important covariates. Under the heterogeneity model, each dataset has twenty important covariates. The three datasets share ten important covariates in common, and the rest important covariates are dataset-specific. Under both models, there are a total of sixty true positives. The nonzero coefficients are randomly generated from a normal distribution with mean zero and variance 0.3125 and 1.25 , representing low and high signal levels. The log event times are generated from the AFT models with intercept equal to 0.5 and $N(0,1)$ random errors. The log censoring times are independently generated from uniform distributions. The overall censoring rate is about 30% .

The simulated data are analyzed using group MCP (GMCP), composite MCP (cMCP), and sparse group MCP (SGMCP). In addition, we also consider two alternatives. The first is a meta-analysis method, where each dataset is analyzed separately using MCP, and then the analysis results are combined across datasets. The second is a pooled analysis method, where the three datasets are combined into a big data matrix, and then variable selection is conducted using MCP. Note that the differences across simulated datasets are smaller than those encountered in practice, which favors meta- and pooled analysis. We acknowledge that multiple other methods are applicable to the simulated data. The two alternatives have the closest framework as the proposed methods.

Summary results based on 200 replicates are shown in Table 1 and 2. Performance of the integrative analysis methods as well as alternatives depend on the similarity of sparsity structures across datasets, correlation structure, and signal level. As an example of the homogeneity model, consider the correlation structure “Banded 2” in Table 1. The homogeneity model favors GMCP, which identifies 34.7 true positives with an average model size 45.2. The cMCP method identifies fewer true positives (30.5). A large number of false positives are identified, with an average model size 149.7. SGMCP identifies 25.6 true positives, with a very small number of false positives (average model size 27.4). In comparison, the meta-analysis and pooled analysis identify much fewer true positives (17.6 and 16.1, respectively). As an example of the heterogeneity model, consider the correlation structure “AR $\rho = 0.5$ ” in Table 2. The cMCP method identifies the most true positives (42.1 on average), but at the price of a large number of false positives (average model size 185.1). GMCP identifies 34.6 true positives. However by forcing the same sparsity structure across datasets, it also identifies a considerable number of false positives (average model size 61.0). SGMCP identifies 26.9 true positives with an average model size 30.2. The meta-analysis and pooled analysis methods identify fewer true positives.

5.3 Analysis of lung cancer prognosis data

In the U.S., lung cancer is the most common cause of cancer death for both men and women. To identify genetic markers associated with the prognosis of lung cancer, gene profiling studies have been extensively conducted. We follow Xie et al. (2011) and collect data from four independent studies with gene expression measurements. The UM (University of Michigan Cancer Center) dataset has a total of 92 patients, with 48 deaths during follow-up. The median follow-up is 55 months. The HLM (Moffitt Cancer Center) dataset has a total of 79 patients, with 60 deaths during follow-up. The median follow-up is 39 months. The DFCI (Dana-Farber Cancer Institute) dataset has a total of 78 patients, with 35 deaths during follow-up. The median follow-up is 51 months. The MSKCC dataset has a total of 102 patients, with 38 deaths during follow-up. The median follow-up is 43.5 months.

Gene expressions were measured using Affymetrix U122 plus 2.0 arrays. A total of 22,283 probe sets were profiled in all four datasets. We first conduct gene expression normalization for each dataset separately, and then normalization across datasets is also conducted to enhance comparability. To further remove noises and improve stability, we conduct a marginal screening and keep the top 2,000 genes for downstream analysis. The expression of each gene in each dataset is normalized to have zero mean and unit variance.

We analyze data using cMCP (Table 3), SGMCP (Table S2.1), meta-analysis (Table S2.2), pooled analysis (Table S2.3), and GMCP (Table S2.4). Although there is overlap, different methods identify significantly different sets of genes. The cMCP method identifies more genes, particularly many more than SGMCP. Such a result fits the pattern observed in simulation. Unlike in simulation, we are not able to objectively evaluate the marker selection results. To provide further insights, we evaluate prediction performance using a cross-validation based approach. Specifically, we split the samples into a training and a testing set with size 3:1. Estimates are generated using the training set samples and used to make prediction for the testing set samples. We separate the testing set samples into two sets with equal sizes based on $\mathbf{X}^m \boldsymbol{\beta}^m$'s. The logrank statistic is computed, evaluating survival difference of the two sets. To reduce the risk of an extreme split, we repeat this process 100 times and compute the average logrank statistics as 7.65 (cMCP), 4.95 (SGMCP), 5.35 (meta-analysis), 5.2 (pooled analysis), and 6.45 (GMCP). All methods are able to separate samples into sets with different survival risk. The cMCP method has the best prediction performance (p-value 0.0057).

6 Discussion

In this article, we have studied the integrative analysis of survival data under the AFT model. The existing research on this topic has been scattered, and this study is the first to systematically study this complicated problem. Both the homogeneity and heterogeneity models have been considered, along with multiple penalization methods. Significantly advancing from the existing studies, the present study rigorously establishes the selection and estimation consistency properties. Although some theoretical development has been

motivated by the existing studies, the heterogeneity across multiple datasets and specific data and model settings make this study unique. Especially, the properties of sparse group penalization have not been studied in single-dataset analysis. Thus this study has both methodological and theoretical contributions. The computational aspect is similar to that in the literature and is largely omitted. Tuning parameter selection using cross validation shows reasonable performance in simulation and data analysis. Theoretical investigation on the consistency of cross validation is very much challenging and postponed. Another contribution is that this study directly compares different methods. The advantage of GMCP under the homogeneity model is expected. Under the heterogeneity model, cMCP may identify a few more true positives, however, at the price of a large number of false positives. The theoretical study does not provide an explanation to this observation. More studies on finite sample properties are needed. In simulation, a total of 24 settings are considered and show similar patterns. More extensive simulations may be pursued in the future. In data analysis, different methods identify different sets of genes. The observed patterns are similar to those in simulation. In addition, cMCP identifies the most genes but also has the best prediction performance. More extensive, especially biological studies may be needed to fully comprehend the data analysis results. In this study, we have focused on survival data and the AFT model. Extensions to other data and model are of interest to future study.

References

- Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface* **2**, 369–380.
- Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *Technical Report, Stanford University*.
- Guerra, R. and Goldstein, D.R. (2009). *Meta-Analysis and Combining Information in Genetics and Genomics*. Chapman and Hall/CRC.
- Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models, *Statistical Science*, 27, 481–499.

- Liu, J., Huang, J., Xie, Y., and Ma, S. (2013). Sparse group penalized integrative analysis of multiple cancer prognosis datasets. *Genetics research* **95**, 68–77.
- Liu, J., Huang, J., and Ma, S. (2014). Integrative analysis of cancer diagnosis studies with composite penalization, *Scandinavian Journal of Statistics* **41**, 87–103.
- Ma, S., Huang, J., and Song, X. (2011a). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* **12**, 763–775.
- Ma, S., Huang, J., Wei, F., Xie, Y., and Fang, K. (2011b). Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in medicine* **30**, 3361–3371.
- Ma, S., Dai, Y., Huang, J., and Xie, Y. (2012). Identification of breast cancer prognosis markers via integrative analysis. *Computational statistics & data analysis* **56**, 2718–2728.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis* **45**, 89–103.
- Xie, Y., Xiao, G., Coombes, K., Behrens, C., Solis, L., Raso, G., Girard, L., Erickson, H., Roth, J., Heymach, J., Moran, C., Danenberg, K., Minna, J., and Wistuba, I. (2011). Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin Cancer Res.* **17**, 5705–5714.
- Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21**, 76–99.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.
- Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7**, 2541–2563.

Appendix

This file contains proofs (Section S1) for the theoretical results described in the main text as well as additional numerical results (Section S2).

S1 Proofs

Let

$$\mathbf{y}^m = W_m^{1/2} \mathbf{Y}^m \text{ and } X^m = W_m^{1/2} \mathbf{X}^m. \quad (\text{S1.1})$$

Then $(\mathbf{Y}^m - \mathbf{X}^m \boldsymbol{\beta}^m)^\top W_m (\mathbf{Y}^m - \mathbf{X}^m \boldsymbol{\beta}^m)$ can be rewritten as $\|\mathbf{y}^m - X^m \boldsymbol{\beta}^m\|^2$, where $\|\cdot\|$ is the ℓ_2 norm. Moreover, we can easily see that

$$\mathbf{y}^m = X^m \boldsymbol{\beta}^m + W_m^{1/2} \boldsymbol{\epsilon}^m. \quad (\text{S1.2})$$

Proof of Theorem 1. First, we prove that

$$\Pr \left\{ \|\hat{\boldsymbol{\beta}}_S^m - \boldsymbol{\beta}_S^{m*}\|_2 < \lambda \frac{4}{\underline{\rho}_1^m} \frac{n}{n_m}, m = 1, \dots, M \right\} \geq 1 - \tau_1,$$

where $\tau_1 = \sum_{m=1}^M \exp \left(-\frac{\lambda^2 n^2}{2\sigma_m^2 \bar{\rho}_2^m n_m} \right)$. Recall that $\hat{\boldsymbol{\beta}}_B = \arg \min_{\boldsymbol{\beta}_B} L(\boldsymbol{\beta}_B)$, where

$$L(\boldsymbol{\beta}_B) = \frac{1}{2n} \sum_{m=1}^M \|\mathbf{y}^m - X_S^m \boldsymbol{\beta}_S^m\|^2 + \lambda \sum_{j \in S} \|\boldsymbol{\beta}_j\|_2.$$

Let $r_m = \lambda \sqrt{|S|} \frac{4}{\underline{\rho}_1^m} \frac{n}{n_m}$ and $\mathfrak{I} = \{\boldsymbol{\beta}_B : \|\boldsymbol{\beta}_S^m - \boldsymbol{\beta}_S^{m*}\|_2 = r_m, m = 1, \dots, M\}$. It suffices to show that

$$\Pr \left(\inf_{\boldsymbol{\beta}_B \in \mathfrak{I}} L(\boldsymbol{\beta}_B) > L(\boldsymbol{\beta}_B^*) \right) \geq 1 - \tau_1.$$

This implies that with probability at least $1 - \tau_1$, $L(\boldsymbol{\beta}_B)$ has a local minimum $\hat{\boldsymbol{\beta}}_B$ that satisfies $\|\hat{\boldsymbol{\beta}}_S^m - \boldsymbol{\beta}_S^{m*}\|_2 < \lambda \sqrt{|S|} \frac{4}{\underline{\rho}_1^m} \frac{n}{n_m}$, for $m = 1, \dots, M$.

Let $\mathbf{u} \in R^{p \times M}$ with $\|\mathbf{u}_S^m\|_2 = 1$, $m = 1, \dots, M$. Define $\boldsymbol{\beta}_S^m = \boldsymbol{\beta}_S^{m*} + r_m \mathbf{u}_S^m$. Consider $Q(\mathbf{u}_B) = n \{L(\boldsymbol{\beta}_B) - L(\boldsymbol{\beta}_B^*)\}$. Obviously, it is equivalent to show that

$$\Pr \left(\inf_{\|\mathbf{u}^m\|_2=1, m=1, \dots, M} Q(\mathbf{u}_B) > 0 \right) \geq 1 - \tau_1. \quad (\text{S1.3})$$

Table 1: Simulation at the low signal level. In each cell, the first row is the number of true positives (sd), and the second row is the number of model size (sd).

Correlation	Meta	Pooled	GMCP	cMCP	SGMCP
Homogeneity model					
AR $\rho = 0.2$	30.3(5.7)	29.0(8.4)	48.8(6.2)	42.6(4.2)	36.5(6.7)
	62.4(19.1)	56.5(29.3)	57.4(9.4)	193.2(13.9)	39.1(8.3)
AR $\rho = 0.5$	20.4(6.0)	18.3(6.7)	39.5(7.9)	33.3(8.1)	28.6(6.9)
	38.7(17.7)	31.2(16.3)	50.8(12.4)	160.6(83.0)	30.9(9.1)
AR $\rho = 0.8$	10.9(2.6)	10.3(3.3)	24.8(7.7)	18.3(4.1)	16.8(5.2)
	17.9(6.1)	15.5(6.4)	34.4(12.8)	75.4(59.2)	18.6(7.2)
Banded 1	26.7(5.8)	25.1(7.6)	46.2(7.6)	40.3(4.5)	34.7(6.2)
	54.3(18.7)	48.7(26.1)	56.5(12.6)	196.6(12.7)	37.8(8.9)
Banded 2	17.6(4.5)	16.1(5.0)	34.7(8.3)	30.5(6.0)	25.6(5.9)
	30.4(11.6)	25.4(12.5)	45.2(13.7)	149.7(95.0)	27.4(7.2)
Banded 3	17.7(5.3)	16.2(4.9)	37.3(7.3)	31.4(5.8)	26.1(6.3)
	32.1(18.6)	26.8(12.9)	51.1(13.7)	166.3(81.7)	28.2(7.6)
Heterogeneity model					
AR $\rho = 0.2$	21.3(5.1)	20.2(5.7)	26.0(9.0)	37.6(5.2)	22.5(7.2)
	35.5(13.8)	31.4(13.9)	53.0(20.3)	199.2(40.3)	28.4(11.0)
AR $\rho = 0.5$	16.8(5.1)	16.7(5.3)	22.8(6.2)	31.7(6.9)	18.8(5.7)
	28.5(10.8)	27.3(12.0)	45.5(15.2)	154.8(94.4)	21.9(7.7)
AR $\rho = 0.8$	10.6(3.8)	10.3(3.5)	15.2(5.5)	20.0(4.9)	11.9(4.2)
	17.0(6.3)	15.3(6.3)	31.4(12.9)	99.9(84.4)	15.3(6.8)
Banded 1	20.4(4.8)	19.9(6.0)	25.2(6.7)	35.3(6.7)	20.9(6.0)
	35.2(15.2)	31.3(13.9)	48.9(14.5)	172.2(77.9)	24.9(7.9)
Banded 2	16.1(4.0)	15.1(3.9)	21.4(6.1)	28.0(5.4)	17.5(4.8)
	24.9(8.4)	22.8(7.7)	44.0(12.2)	129.9(103.4)	21.0(6.2)
Banded 3	15.9(3.6)	15.2(4.4)	20.2(6.0)	27.1(6.2)	17.8(4.9)
	26.8(10.8)	24.3(10.2)	43.3(14.2)	102.7(115.7)	22.3(7.5)

Table 2: Simulation at the high signal level. In each cell, the first row is the number of true positives (sd), and the second row is the number of model size (sd).

Correlation	Meta	Pooled	GMCP	cMCP	SGMCP
Homogeneity model					
AR $\rho = 0.2$	39.4(4.5)	39.2(5.4)	58.3(2.3)	52.3(2.9)	49.9(3.8)
	49.9(9.3)	48.8(11.4)	60.1(4.2)	174.6(11.6)	50.1(4.1)
AR $\rho = 0.5$	30.1(5.0)	30.0(6.0)	55.4(3.6)	46.5(3.3)	44.2(4.0)
	42.0(10.1)	41.8(12.3)	58.3(4.3)	179.8(15.3)	44.5(4.2)
AR $\rho = 0.8$	17.4(3.8)	17.1(3.9)	46.5(6.7)	29.5(6.4)	29.6(5.9)
	24.2(6.5)	23.6(7.3)	54.1(10.6)	103.8(97.8)	30.7(6.1)
Banded 1	36.9(4.7)	35.9(5.1)	57.2(2.7)	50.3(2.9)	47.9(4.3)
	47.3(8.4)	43.7(7.6)	58.7(4.4)	178.4(12.1)	48.3(4.4)
Banded 2	25.9(4.3)	25.5(4.7)	53.3(4.6)	41.1(3.4)	38.6(5.5)
	36.3(8.8)	34.4(9.1)	57.8(8.3)	186.2(16.7)	39.7(6.1)
Banded 3	27.1(3.8)	26.5(4.3)	53.7(4.5)	42.4(4.4)	40.8(4.7)
	37.3(8.4)	35.8(8.3)	57.8(7.0)	179.8(21.4)	42.0(5.6)
Heterogeneity model					
AR $\rho = 0.2$	34.4(4.1)	34.0(4.1)	40.0(4.2)	48.91(3.2)	33.9(4.6)
	39.7(6.0)	37.9(4.8)	69.2(7.9)	180.4(18.9)	36.6(4.7)
AR $\rho = 0.5$	25.9(4.5)	24.1(5.9)	34.6(5.7)	42.1(4.1)	26.9(4.8)
	32.7(6.6)	29.5(7.3)	61.0(9.8)	185.1(18.0)	30.2(6.2)
AR $\rho = 0.8$	16.4(3.4)	15.6(3.5)	23.7(5.6)	26.8(5.3)	17.5(4.4)
	22.2(5.1)	21.3(6.5)	44.3(10.3)	157.5(87.3)	20.9(5.6)
Banded 1	30.8(4.1)	30.2(4.6)	36.8(5.3)	45.8(3.1)	30.0(5.2)
	36.0(5.8)	35.4(6.7)	64.1(9.3)	177.7(17.3)	32.6(6.7)
Banded 2	22.9(4.6)	22.4(4.1)	32.1(5.9)	36.6(4.3)	25.2(4.9)
	29.3(7.8)	27.5(5.4)	57.4(8.4)	169.2(51.2)	28.6(5.3)
Banded 3	23.0(4.6)	22.6(4.2)	31.6(6.2)	37.4(5.0)	24.2(6.8)
	28.7(5.8)	27.9(5.3)	57.1(9.9)	169.2(42.1)	26.6(7.5)

Table 3: Analysis of lung cancer data using cMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	DFCI	MSKCC
201462_at	SCRN1		0.0045		
202637_s_at	ICAM1				0.0037
203240_at	FCGBP		0.0024		
203876_s_at	MMP11				-0.0013
203917_at	CXADR				0.0040
203921_at	CHST2	0.0024			
204855_at	SERPINB5	-0.0008			
205234_at	SLC16A4				-0.0016
205399_at	DCLK1				-0.0031
206461_x_at	MT1H			-0.0008	
206754_s_at	CYP2B6			0.0048	
206994_at	CST4		-0.0017		
207850_at	CXCL3		-0.0155		
208025_s_at	HMGA2			-0.0016	
208451_s_at	C4A		0.0038		
208607_s_at	SAA2				0.0044
209343_at	EFHD1				0.0028
212328_at	LIMCH1			0.0028	
212338_at	MYO1D		0.0019		
213338_at	TMEM158			-0.0003	
214452_at	BCAT1				0.0004
215867_x_at	CA12		-0.0054		
218677_at	S100A14				-0.0081
219654_at	PTPLA		-0.0109		
219747_at	NDNF	0.0001			
220952_s_at	PLEKHA5		-0.0018		
221841_s_at	KLF4	-0.0024			
222043_at	CLU		0.0008		

Together with (S1.1) and (S1.2), we have

$$\begin{aligned}
Q(\mathbf{u}_B) &= \frac{1}{2} \sum_{m=1}^M (\|\mathbf{y}^m - X_S^m (\boldsymbol{\beta}_S^{m*} + r_m \mathbf{u}_S^m)\|^2 - \|\mathbf{y}^m - X_S^m \boldsymbol{\beta}_S^{m*}\|^2) \\
&\quad + n\lambda \sum_{j \in S} \left\{ \|\boldsymbol{\beta}_j^* + \mathbf{r} \circ \mathbf{u}_j\|_2 - \|\boldsymbol{\beta}_j^*\|_2 \right\} \\
&= - \sum_{m=1}^M r_m \mathbf{u}_S^{m\top} \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m + \frac{1}{2} \sum_{m=1}^M r_m^2 \mathbf{u}_S^{m\top} \mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m \mathbf{u}_S^m \\
&\quad + n\lambda \sum_{j \in S} \left\{ \|\boldsymbol{\beta}_j^* + \mathbf{r} \circ \mathbf{u}_j\|_2 - \|\boldsymbol{\beta}_j^*\|_2 \right\} \\
&=: Q_1 + Q_2 + Q_3,
\end{aligned} \tag{S1.4}$$

where $\mathbf{r} = (r_1, \dots, r_M)^\top$, and \circ denotes the Hadamard (component-wise) product. Write $Q_1 = \sum_{m=1}^M Q_{1m}$ where $Q_{1m} = -r_m \mathbf{u}_S^{m\top} \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m$. Note that $\|W_m \mathbf{X}_S^m \mathbf{u}_S^m\|_2^2 \leq n_m \bar{\rho}_2^m$. With the sub-Gaussian tail as specified in Condition 1, we have for any given ε_m

$$\Pr(|Q_{1m}| > r_m \varepsilon_m) \leq 2 \exp \left(-\frac{\varepsilon_m^2}{2\sigma_m^2 \|W_m \mathbf{X}_S^m \mathbf{u}_S^m\|_2^2} \right) \leq 2 \exp \left(-\frac{\varepsilon_m^2}{2n_m \bar{\rho}_2^m \sigma_m^2} \right).$$

Together with the Bonferroni's inequality, we have

$$\Pr(Q_1 < -\sum_{m=1}^M r_m \varepsilon_m) \leq \sum_{m=1}^M \Pr(Q_{1m} < -r_m \varepsilon_m) \leq \sum_{m=1}^M \exp \left(-\frac{\varepsilon_m^2}{2n_m \bar{\rho}_2^m \sigma_m^2} \right).$$

Set $\varepsilon_m = \frac{1}{4} \underline{\rho}_1^m n_m r_m$. Then

$$\Pr(Q_1 \geq -\frac{1}{4} \sum_{m=1}^M r_m^2 n_m \underline{\rho}_1^m) \geq 1 - \sum_{m=1}^M \exp \left(-\frac{n_m r_m^2 (\underline{\rho}_1^m)^2}{32 \bar{\rho}_2^m \sigma_m^2} \right). \tag{S1.5}$$

For Q_2 , since $\mathbf{u}_S^{m\top} \mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m \mathbf{u}_S^m \geq n_m \underline{\rho}_1^m$, we have

$$Q_2 \geq \frac{1}{2} \sum_{m=1}^M r_m^2 n_m \underline{\rho}_1^m. \tag{S1.6}$$

Term Q_3 can be dealt with as follows. By the Triangle inequality and $(\sum_{i=1}^d |v_i|)^2 \leq d \sum_{i=1}^d v_i^2$,

for any sequence v_i , we have

$$\begin{aligned} \sum_{j \in S} \|\beta_j^* + \mathbf{r} \circ \mathbf{u}_j\|_2 - \|\beta_j^*\|_2 &\leq \sum_{j \in S} \|\mathbf{r} \circ \mathbf{u}_j\|_2 \\ &\leq \sqrt{|S|} \sqrt{\sum_{j \in S} \|\mathbf{r} \circ \mathbf{u}_j\|_2^2} = \sqrt{|S|} \sqrt{\sum_{m=1}^M r_m^2} \leq \sqrt{|S|} \sum_{m=1}^M r_m. \end{aligned}$$

Therefore, we have that term Q_3 satisfies

$$|Q_3| \leq n\lambda \sqrt{|S|} \sum_{m=1}^M r_m. \quad (\text{S1.7})$$

Combining (S1.4), (S1.5), (S1.6), and (S1.7), we have

$$Q(\mathbf{u}_S) \geq \frac{1}{4} \sum_{m=1}^M r_m^2 n_m \rho_1^m - n\lambda \sqrt{|S|} \sum_{m=1}^M r_m := L(\mathbf{r}) \quad (\text{S1.8})$$

with probability at least $1 - \sum_{m=1}^M \exp\left(-\frac{n_m r_m^2 (\rho_1^m)^2}{32 \rho_2^m \sigma_m^2}\right)$. Recall that $r_m = \lambda \sqrt{|S|} \frac{4}{\rho_1^m} \frac{n}{n_m}$. Then $L(\mathbf{r}) > 0$ with probability at least $1 - \sum_{m=1}^M \exp\left(-\frac{\lambda^2 |S| n^2}{2 \sigma_m^2 \rho_2^m n_m}\right)$. Therefore, (S1.3) is proved, and Part 1 of Theorem 1 is established.

Now consider Part 2. By the Karush-Kuhn-Tucker (KKT) conditions, we need to prove that for $m = 1, \dots, M$,

$$-X_S^{m\top} (\mathbf{y}^m - X_S^m \hat{\beta}_S^m) + n\lambda \frac{\hat{\beta}_S^m}{\|\hat{\beta}_B\|_2} = 0, \quad (\text{S1.9})$$

$$\|X_{S^c}^\top (\mathbf{y}^m - X_S^m \hat{\beta}_S^m)\|_\infty \leq n\lambda. \quad (\text{S1.10})$$

Then $\hat{\beta}^{glasso} = \{\hat{\beta}_B^{glasso}, \hat{\beta}_{B^c}^{glasso}\}$ with $\hat{\beta}_B^{glasso} = \hat{\beta}_B$, $\hat{\beta}_{B^c}^{glasso} = 0$ is a local minimizer of (3). From Part 1, $\tilde{\beta}_S$ minimizes

$$L(\beta_B) = \frac{1}{2n} \sum_{m=1}^M \|\mathbf{y}^m - X_S^m \beta_S^m\|^2 + \lambda \sum_{j \in S} \|\beta_j\|_2.$$

Therefore, (S1.9) holds, together with (S1.2) which also yields

$$\hat{\beta}_S^m - \beta_S^{m*} = (X_S^{m\top} X_S^m)^{-1} \left\{ X_S^{m\top} W_m^{-1/2} \epsilon^m - n\lambda \frac{\hat{\beta}_S^m}{\|\hat{\beta}_B\|_2} \right\}. \quad (\text{S1.11})$$

Note that

$$X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \hat{\boldsymbol{\beta}}_S^m) = X_{S^c}^{m\top} W_m^{1/2} \boldsymbol{\epsilon}^m - X_{S^c}^{m\top} X_S^m (\hat{\boldsymbol{\beta}}_S^m - \boldsymbol{\beta}_S^{m*}). \quad (\text{S1.12})$$

Substituting (S1.11) into (S1.12), we obtain

$$\begin{aligned} & \|X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \hat{\boldsymbol{\beta}}_S^m)\|_\infty \\ &= \left\| X_{S^c}^{m\top} W_m^{1/2} \boldsymbol{\epsilon}^m - X_{S^c}^{m\top} X_S^m (X_S^{m\top} X_S^m)^{-1} \left\{ X_S^{m\top} W_m^{1/2} \boldsymbol{\epsilon}^m - n\lambda \frac{\hat{\boldsymbol{\beta}}_S^m}{\|\hat{\boldsymbol{\beta}}_B\|_2} \right\} \right\|_\infty \\ &\leq \left\| X_{S^c}^{m\top} W_m^{1/2} \boldsymbol{\epsilon}^m \right\|_\infty + \left\| X_{S^c}^{m\top} X_S^m (X_S^{m\top} X_S^m)^{-1} X_S^{m\top} W_m^{1/2} \boldsymbol{\epsilon}^m \right\|_\infty \\ &\quad + n\lambda \left\| X_{S^c}^{m\top} X_S^m (X_S^{m\top} X_S^m)^{-1} \frac{\hat{\boldsymbol{\beta}}_S^m}{\|\hat{\boldsymbol{\beta}}_B\|_2} \right\|_\infty \\ &\leq \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + \left\| \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1} \right\|_\infty \left\| \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \\ &\quad + n\lambda \left\| \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1} \right\|_\infty \left\| \frac{\hat{\boldsymbol{\beta}}_S^m}{\|\hat{\boldsymbol{\beta}}_B\|_2} \right\|_\infty \\ &\leq \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + \psi_m \left\| \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + n\lambda \psi_m \end{aligned} \quad (\text{S1.13})$$

By the condition $\psi_m \leq D_m < 1$, if

$$\left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \leq n\lambda \frac{1 - D_m}{1 + D_m}, \quad (\text{S1.14})$$

then from (S1.13) it follows

$$\begin{aligned} \left\| X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \tilde{\boldsymbol{\beta}}_S^m) \right\|_\infty &\leq \left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty (1 + \psi_m) + n\lambda \psi_m \\ &\leq n\lambda(1 - D_m) + n\lambda D_m = n\lambda. \end{aligned}$$

We now derive the probability bounds for the event in (S1.14). By the Bonferroni's inequality and sub-Gaussian tail probability bound in Condition 1,

$$\begin{aligned} & \Pr \left\{ \left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty > n\lambda \frac{1 - D_m}{1 + D_m}, \text{ for } m = 1, \dots, M \right\} \\ &\leq p \sum_{m=1}^M \Pr \left\{ |\mathbf{X}_j^{m\top} W_m \boldsymbol{\epsilon}^m| > n\lambda \frac{1 - D_m}{1 + D_m} \right\} \\ &\leq 2p \sum_{m=1}^M \exp \left\{ -\frac{n^2 \lambda^2 (1 - D_m)^2}{2n_m \Lambda_m \sigma_m^2 (1 + D_m)^2} \right\}. \end{aligned} \quad (\text{S1.15})$$

Then Part 2 is established by combining Part1, (S1.9), (S1.10), and (S1.15). \square

Proof of Theorem 2. Recall that $\tilde{\beta}_{\mathcal{B}} = \arg \min_{\beta_{\mathcal{B}}} H(\beta_{\mathcal{B}})$, where

$$H(\beta_{\mathcal{B}}) = \frac{1}{2n} \sum_{m=1}^M \|\mathbf{y}^m - \mathbf{X}_S^m \beta_S^m\|^2.$$

Let $r_m = \sqrt{\frac{|S|}{n}} R_m$ with $R_m \in (0, \infty)$ and $\mathfrak{I} = \{\beta_{\mathcal{B}} : \|\beta_S^m - \beta_S^{m*}\|_2 = r_m, m = 1, \dots, M\}$. Similar as the proof of part 1 in Theorem 1, if we can prove

$$\Pr \left(\inf_{\beta_{\mathcal{B}} \in \mathfrak{I}} H(\beta_{\mathcal{B}}) > H(\beta_{\mathcal{B}}^*) \right) \geq 1 - \sum_{m=1}^M \exp \left\{ -R_m^2 \frac{|S|(\rho_1^m)^2}{8\bar{\rho}_2^m \sigma_m^2} \right\}, \quad (\text{S1.16})$$

then $H(\beta_{\mathcal{B}})$ has a local minimum $\hat{\beta}_{\mathcal{B}}$ that satisfies $\|\hat{\beta}_S^m - \beta_S^{m*}\|_2 < r_m, m = 1, \dots, M$ with probability at least $1 - \sum_{m=1}^M \exp \left\{ -R_m^2 \frac{|S|(\rho_1^m)^2}{8\bar{\rho}_2^m \sigma_m^2} \right\}$.

Together with (S1.1) and (S1.2), we have

$$\begin{aligned} H(\beta_{\mathcal{B}}) - H(\beta_{\mathcal{B}}^*) &= - \sum_{m=1}^M (\hat{\beta}_S^m - \beta_S^{m*})^\top \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \\ &\quad + \frac{1}{2} \sum_{m=1}^M (\hat{\beta}_S^m - \beta_S^{m*})^\top \mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m (\hat{\beta}_S^m - \beta_S^{m*}) \\ &=: H_1 + H_2, \end{aligned} \quad (\text{S1.17})$$

For H_2 , since $\lambda_{\min} \{n_m^{-1} \mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m\} = \rho_1^m$ and $\|\beta_S^m - \beta_S^{m*}\|_2 = r_m$, we have

$$H_2 \geq \frac{1}{2} \sum_{m=1}^M r_m^2 n_m \rho_1^m. \quad (\text{S1.18})$$

For H_1 we have for any ε_m ,

$$\begin{aligned} \Pr(H_1 \leq - \sum_{m=1}^M r_m \varepsilon_m) &\leq \sum_{m=1}^M \exp \left(- \frac{r_m^2 \varepsilon_m^2}{2\sigma_m^2 \|\mathbf{X}_S^m (\hat{\beta}_S^m - \beta_S^{m*})\|_2^2} \right) \\ &\leq \sum_{m=1}^M \exp \left(- \frac{\varepsilon_m^2}{2n_m \bar{\rho}_2^m \sigma_m^2} \right). \end{aligned}$$

The first inequality holds due to the sub-Gaussian tail probability under Condition 1, and the last inequality holds due to the fact that $\|\mathbf{X}_S^m (\hat{\beta}_S^m - \beta_S^{m*})\|_2^2 \leq n_m \bar{\rho}_2^m r_m^2$. Set $\varepsilon_m =$

$\frac{1}{2}\rho_1^m n_m r_m$. Then

$$\Pr(H_1 > -\frac{1}{2} \sum_{m=1}^M r_m^2 n_m \rho_1^m) \geq 1 - \sum_{m=1}^M \exp\left(-\frac{n_m r_m^2 (\rho_1^m)^2}{8\bar{\rho}_2^m \sigma_m^2}\right). \quad (\text{S1.19})$$

Recall that $r_m = \sqrt{\frac{|S|}{n}} R_m$. Combining (S1.17), (S1.18) and (S1.19), we have (S1.16) holds. This complete the proof of Part 1.

Next, we prove Part 2. By the Karush-Kuhn-Tucher(KKT) conditions, we need to prove that $\hat{\beta}^{oracle}$ satisfies

$$-X_S^{m\top} (\mathbf{y}^m - X_S^m \tilde{\beta}_S^m) + np'_\lambda(\|\tilde{\beta}_B\|_2) \circ \frac{\tilde{\beta}_S^m}{\|\tilde{\beta}_B\|_2} = 0, \quad (\text{S1.20})$$

$$\|X_{S^c}^{m\top} (\mathbf{y}^m - X_S^m \tilde{\beta}_S^m)\|_\infty \leq np'_\lambda(0+). \quad (\text{S1.21})$$

If $\min_{j \in S} \|\tilde{\beta}_j\|_2 > \theta\lambda$, $p'_\lambda(\|\tilde{\beta}_B\|_2) = 0$, and certainly (S1.20) holds. Define

$$R_m^\dagger \leq \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\sqrt{M}} \sqrt{\frac{n_m}{|S|}}.$$

Note that $\lambda < \frac{\min_{j \in S} \|\beta_j^*\|_2}{2\theta}$. Therefore, we can conclude the event

$$\left\{ \|\tilde{\beta}_S^m - \beta_S^{m*}\|_2 \leq \sqrt{\frac{|S|}{n_m}} R_m^\dagger, \quad m = 1, \dots, M \right\}$$

belongs to the event $\left\{ \min_{j \in S} \|\tilde{\beta}_j\|_2 > \theta\lambda \right\}$. That is,

$$\begin{aligned} \Pr\left\{ \min_{j \in S} \|\tilde{\beta}_j\|_2 > \theta\lambda \right\} &\geq \Pr\left(\|\tilde{\beta}_S^m - \beta_S^{m*}\|_2 \leq \sqrt{\frac{|S|}{n_m}} R_m^\dagger, \quad m = 1, \dots, M \right) \\ &\geq 1 - \sum_{m=1}^M \exp\left\{ -\frac{|S|(\rho_1^m)^2}{8\sigma_m^2 \bar{\rho}_2^m} R_m^{\dagger 2} \right\}. \end{aligned} \quad (\text{S1.22})$$

Now consider the probability of

$$\|X_{S^c}^{m\top} (\mathbf{y}^m - X_S^m \tilde{\beta}_S^m)\|_\infty \leq np'_\lambda(0+), \quad \text{for } m = 1, \dots, M. \quad (\text{S1.23})$$

Note that

$$X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \tilde{\boldsymbol{\beta}}_S^m) = \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m - \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\tilde{\boldsymbol{\beta}}_S^m - \boldsymbol{\beta}_S^{m*}). \quad (\text{S1.24})$$

Combining (S1.23) and (S1.24), we can obtain

$$\begin{aligned} & \|X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \tilde{\boldsymbol{\beta}}_S^m)\|_\infty \\ = & \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m - \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1} \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \\ \leq & \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + \left\| \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1} \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \\ \leq & \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + \left\| \mathbf{X}_{S^c}^{m\top} W_m \mathbf{X}_S^m (\mathbf{X}_S^{m\top} W_m \mathbf{X}_S^m)^{-1} \right\|_\infty \left\| \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \\ \leq & \left\| \mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty + \psi_m \left\| \mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \\ \leq & \left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty (1 + \psi_m). \end{aligned} \quad (\text{S1.25})$$

If

$$\left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty \leq \frac{np'_\lambda(0+)}{(1 + \psi_m)}, \quad (\text{S1.26})$$

then from (S1.25) it follows

$$\|X_{S^c}^{m\top}(\mathbf{y}^m - X_S^m \tilde{\boldsymbol{\beta}}_S^m)\|_\infty \leq \frac{np'_\lambda(0+)}{(1 + \psi_m)}(1 + \psi_m) \leq np'_\lambda(0+),$$

which proves (S1.21). We now derive the probability bounds for the event in (S1.26). In fact, by the Bonferroni's inequality and sub-Gaussian tail probability bound under Condition 1,

$$\begin{aligned} & \Pr \left\{ \left\| \mathbf{X}^{m\top} W_m \boldsymbol{\epsilon}^m \right\|_\infty > \frac{np'_\lambda(0+)}{(1 + \psi_m)}, \exists m \in \{1, \dots, M\} \right\} \\ \leq & p \sum_{m=1}^M \Pr \left\{ \left| \mathbf{X}_j^{m\top} W_m \boldsymbol{\epsilon}^m \right| > \frac{np'_\lambda(0+)}{(1 + \psi_m)} \right\} \\ \leq & 2p \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_\lambda'^2(0+)}{2n_m \Lambda_m \sigma_m^2 (1 + \psi_m)^2} \right\}. \end{aligned} \quad (\text{S1.27})$$

Part (2) is proved by combining (S1.20), (S1.21), (S1.22), and (S1.27). \square

Proof of Theorem 3. The proof is similar to that of Part 1 of Theorem 2 and is omitted here. \square

Proof of Theorem 4. By the Karush-Kuhn-Tucher(KKT) conditions, we need to prove that $\check{\beta}$ satisfies

$$-X_{S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\beta}_{S_m}^m) + np'_{O,\lambda_O}(\sum_{m=1}^M p_{I,\lambda_I}(|\check{\beta}_{S_m}^m|)) \circ p'_{I,\lambda_I}(|\check{\beta}_{S_m}^m|) = 0, \quad (\text{S1.28})$$

$$|X_{S-S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\beta}_{S_m}^m)| \leq np'_{I,\lambda_I}(0+)p'_{O,\lambda_O}(\sum_{m=1}^M p_{I,\lambda_I}(|\check{\beta}_{S-S_m}^m|)), \quad (\text{S1.29})$$

$$\|X_{S^c}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\beta}_{S_m}^m)\|_\infty \leq np'_{O,\lambda_O}(0+)p'_{I,\lambda_I}(0+). \quad (\text{S1.30})$$

If $\min_{j \in S_m} |\check{\beta}_j^m| > \theta_I \lambda_I$, then $p'_{I,\lambda_I}(|\check{\beta}_{S_m}^m|) = 0$. Recall the definition of the estimator $\check{\beta}_{S_m}^m$. We can easily get (S1.28). Set

$$C_m^\dagger \leq \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2} \sqrt{\frac{n_m}{|S_m|}}.$$

Note that $\lambda_I < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_I}$. Therefore,

$$\begin{aligned} \Pr \left\{ \min_{(j,m) \in \mathcal{A}} |\check{\beta}_j^m| > \theta_I \lambda_I \right\} &\geq \Pr \left(\|\check{\beta}_{S_m}^m - \beta_{S_m}^{m*}\|_2 \leq \sqrt{\frac{|S_m|}{n_m}} C_m^\dagger, \quad m = 1, \dots, M \right) \\ &\geq 1 - 2 \sum_{m=1}^M \exp \left\{ -C_m^{\dagger 2} \frac{|S_m|(\rho_1^{*m})^2}{8\bar{\rho}_2^{*m}\sigma_m^2} \right\}. \end{aligned} \quad (\text{S1.31})$$

In fact,

$$X_{S-S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\beta}_{S_m}^m) = \mathbf{X}_{S-S_m}^{m\top} W_m \boldsymbol{\epsilon}^m - \mathbf{X}_{S-S_m}^{m\top} W_m X_{S_m}^m (\check{\beta}_{S_m}^m - \beta_{S_m}^{m*}),$$

and $\check{\beta}_{S_m}^m - \beta_{S_m}^{m*} = (\mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m)^{-1} X_{S_m}^{m\top} W_m \boldsymbol{\epsilon}^m$. Then we have

$$\begin{aligned} &|X_{S-S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\beta}_{S_m}^m)| \\ &\leq |\mathbf{X}_{S-S_m}^{m\top} W_m \boldsymbol{\epsilon}^m| + |\mathbf{X}_{S-S_m}^{m\top} W_m X_{S_m}^m (\mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m)^{-1} X_{S_m}^{m\top} W_m \boldsymbol{\epsilon}^m| \\ &\leq |\mathbf{X}_{S-S_m}^{m\top} W_m \boldsymbol{\epsilon}^m| + \left\| \mathbf{X}_{S-S_m}^{m\top} W_m X_{S_m}^m (\mathbf{X}_{S_m}^{m\top} W_m \mathbf{X}_{S_m}^m)^{-1} \right\|_\infty |\mathbf{X}_{S_m}^{m\top} W_m \boldsymbol{\epsilon}^m| \\ &\leq \|\mathbf{X}_{S-S_m}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty + \psi_m^* \|\mathbf{X}_{S_m}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty \\ &\leq \|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty (1 + \psi_m^*). \end{aligned} \quad (\text{S1.32})$$

Hence (S1.29) holds when

$$\|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty \leq \frac{np'_{I,\lambda_I}(0+)p'_{O,\lambda_O}(J^{-m} f_I^{max})}{(1 + \psi_m^*)}. \quad (\text{S1.33})$$

That is because for $m = 1, \dots, M$,

$$\begin{aligned}
& |X_{S-S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m)| \leq \frac{np'_{I,\lambda_I}(0+)p'_{O,\lambda_O}(J^{-m}f_I^{max})}{(1+\psi_m^*)}(1+\psi_m^*) \\
& \leq np'_{I,\lambda_I}(0+)p'_{O,\lambda_O}(J^{-m}f_I^{max})(1+\psi_m^*) \\
& \leq np'_{I,\lambda_I}(0+)p'_{O,\lambda_O} \left(\sum_{m=1}^M p_{I,\lambda_I}(|\check{\boldsymbol{\beta}}_{S-S_m}^m|) \right).
\end{aligned}$$

We now derive the probability bounds for the event in (S1.33). In fact, by Bonferroni's inequality and sub-Gaussian tail probability bound in Condition 1,

$$\begin{aligned}
& \Pr \left\{ \|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty > \frac{np'_{I,\lambda_I}(0+)p'_{O,\lambda_O}(J^{-m}f_I^{max})}{(1+\psi_m^*)}, \exists m \in \{1, \dots, M\} \right\} \\
& \leq 2|S| \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{I,\lambda_I}^2(0+)p_{O,\lambda_O}^2(J^{-m}f_I^{max})}{2n_m \bar{\rho}_2^{*m} \sigma_m^2 (1+\psi_m^*)^2} \right\}. \tag{S1.34}
\end{aligned}$$

Similarly, we can prove (S1.30). Actually,

$$\begin{aligned}
& |X_{S^c}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m)|_\infty \leq \|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty + \psi_m^* \|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty \\
& < \|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty + \frac{\psi_m^*}{(1+\psi_m^*)} np'_{O,\lambda_O}(0+)p'_{I,\lambda_I}(0+).
\end{aligned}$$

Based on the above discussions, (S1.30) follows if $\|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty < \frac{np'_{O,\lambda_O}(0+)p'_{I,\lambda_I}(0+)}{(1+\psi_m^*)}$. The probability bound is derived as

$$\begin{aligned}
& \Pr \left\{ \|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty > \frac{np'_{O,\lambda_O}(0+)p'_{I,\lambda_I}(0+)}{(1+\psi_m^*)}, \exists m \in \{1, \dots, M\} \right\} \\
& \leq 2(p - |S|) \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{I,\lambda_I}^2(0+)p_{O,\lambda_O}^2(0+)}{2n_m \Lambda_m \sigma_m^2 (1+\psi_m^*)^2} \right\}. \tag{S1.35}
\end{aligned}$$

Therefore, the theorem is proved by combining (S1.28), (S1.29), (S1.29), (S1.31), (S1.34) and (S1.35). \square

Proof of Theorem 5. By the Karush-Kuhn-Tucher(KKT) conditions, we need to prove that $\check{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned}
& -X_{S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m) + np'_{1,\lambda_1}(\|\check{\boldsymbol{\beta}}_{S_m}\|_2) \circ \frac{\check{\boldsymbol{\beta}}_{S_m}^m}{\|\check{\boldsymbol{\beta}}_{S_m}\|_2} \\
& + np'_{2,\lambda_2}(\|\check{\boldsymbol{\beta}}_{S_m}^m\|) \circ \text{sgn}(\check{\boldsymbol{\beta}}_{S_m}^m) = 0, \tag{S1.36}
\end{aligned}$$

$$\|X_{S-S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m)\|_\infty \leq np'_{2,\lambda_2}(0+), \quad (\text{S1.37})$$

$$\|X_{S^c}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m)\|_\infty \leq np'_{1,\lambda_1}(0+) + np'_{2,\lambda_2}(0+). \quad (\text{S1.38})$$

Note that $\check{\boldsymbol{\beta}}_{S_m}^m$ satisfies $-X_{S_m}^{m\top}(\mathbf{y}^m - X_{S_m}^m \check{\boldsymbol{\beta}}_{S_m}^m) = 0$. If

$$\min_{j \in S_m} |\check{\beta}_j^m| > \theta_2 \lambda_2 \quad \text{and} \quad \min_{j \in S} \|\check{\boldsymbol{\beta}}_j\|_2 > \theta_1 \lambda_1,$$

then we have (S1.36). Set $C_m^\dagger \leq \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2} \sqrt{\frac{n_m}{|S_m|}}$. Note that

$$\lambda_1 < \frac{\min_{j \in S} \|\boldsymbol{\beta}_j^*\|_2}{2\theta_1}, \quad \lambda_2 < \frac{\min_{(j,m) \in \mathcal{A}} |\beta_j^{m*}|}{2\theta_2}.$$

Therefore, if $\min_{(j,m) \in \mathcal{A}} |\check{\beta}_j^m| > \theta_2 \lambda_2$, then we must have $\min_{j \in S} \|\check{\boldsymbol{\beta}}_j\|_2 > \theta_1 \lambda_1$.

$$\begin{aligned} & \Pr \left\{ \min_{(j,m) \in \mathcal{A}} |\check{\beta}_j^m| > \theta_1 \lambda_1, \quad \min_{j \in S} \|\check{\boldsymbol{\beta}}_j\|_2 > \theta_2 \lambda_2 \right\} \\ & \geq \Pr \left(\|\check{\boldsymbol{\beta}}_{S_m}^m - \boldsymbol{\beta}_{S_m}^{m*}\|_2 \leq \sqrt{\frac{|S_m|}{n_m}} C_m^\dagger, \quad m = 1, \dots, M \right) \\ & \geq 1 - 2 \sum_{m=1}^M \exp \left\{ -C_m^{\dagger 2} \frac{|S_m|(\rho_1^{*m})^2}{8\bar{\rho}_2^{*m}\sigma_m^2} \right\}. \end{aligned} \quad (\text{S1.39})$$

Similar as the proof of Theorem 4, (S1.37) holds when

$$\|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty \leq \frac{np'_{2,\lambda_2}(0+)}{(1 + \psi_m^*)}. \quad (\text{S1.40})$$

Then we have

$$\begin{aligned} & \Pr \left\{ \|\mathbf{X}_S^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty > \frac{np'_{2,\lambda_2}(0+)}{(1 + \psi_m^*)}, \quad \exists m \in \{1, \dots, M\} \right\} \\ & \leq 2|S| \sum_{m=1}^M \exp \left\{ -\frac{n^2 p_{2,\lambda_2}'^2(0+)}{2n_m \bar{\rho}_2^{*m} \sigma_m^2 (1 + \psi_m^*)^2} \right\}. \end{aligned} \quad (\text{S1.41})$$

Similarly, we can show (S1.38) holds when $\|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty < \frac{np'_{1,\lambda_1}(0+) + np'_{2,\lambda_2}(0+)}{(1 + \psi_m^*)}$. The probability bound is derived as

$$\Pr \left\{ \|\mathbf{X}_{S^c}^{m\top} W_m \boldsymbol{\epsilon}^m\|_\infty > \frac{np'_{1,\lambda_1}(0+) + np'_{2,\lambda_2}(0+)}{(1 + \psi_m^*)}, \quad \exists m \in \{1, \dots, M\} \right\}$$

$$\leq 2(p - |S|) \sum_{m=1}^M \exp \left\{ -\frac{n^2 [p'_{1,\lambda_1}(0+) + p'_{2,\lambda_2}(0+)]^2}{2n_m \Lambda_m \sigma_m^2 (1 + \psi_m^*)^2} \right\}. \quad (\text{S1.42})$$

Therefore, the theorem is proved by combining (S1.36), (S1.37), (S1.37), (S1.39), (S1.41), and (S1.42). \square

S2 Additional Numerical Results

Table S2.1: Analysis of lung cancer data using SGMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	DFCI	MSKCC
201462_at	SCRN1		0.0034		0.0020
202831_at	GPX2		-0.0022		-0.0021
203917_at	CXADR	0.0021		0.0004	0.0066
205776_at	FMO5	0.0005	0.0035	0.0038	
206754_s_at	CYP2B6	0.0012		0.0020	
207850_at	CXCL3		-0.0216		0.0120
208025_s_at	HMGA2	-0.0028	0.0001	-0.0037	-0.0012
219654_at	PTPLA	-0.0025	-0.0145		0.0055
219764_at	FZD10	-0.0005	-0.0019	-1.6E-05	-0.0022

Table S2.2: Analysis of lung cancer data using meta-analysis: identified genes and their estimates.

Probe	Gene	UM	HLM	DFCI	MSKCC
201462_at	SCRN1		0.0101		
203559_s_at	ABP1		0.0005		
203876_s_at	MMP11				-0.0066
203921_at	CHST2	0.0051			
204855_at	SERPINB5	-0.0012			
206754_s_at	CYP2B6			0.0104	
206994_at	CST4		-0.0037		
207850_at	CXCL3		-0.0246		
208025_s_at	HMGA2	-0.0021		-0.0010	
209343_at	EFHD1				0.0096
212328_at	LIMCH1			0.0050	
213703_at	LINC00342	0.0008			
215867_x_at	CA12		-0.0026		
218677_at	S100A14				-0.0257
218824_at	PNMAL1	0.0003			
219654_at	PTPLA		-0.0240		
219747_at	NDNF	0.0002			
220952_s_at	PLEKHA5		-0.0047		
221841_s_at	KLF4	-0.0047			
222043_at	CLU		0.0049		

Table S2.3: Analysis of lung cancer data using pooled analysis: identified genes and their estimates.

Probe	Gene	UM	HLM	DFCI	MSKCC
201462_at	SCRN1		0.0101		
203559_s_at	ABP1		0.0005		
203876_s_at	MMP11				-0.0066
203921_at	CHST2	0.0051			
204855_at	SERPINB5	-0.0012			
206754_s_at	CYP2B6			0.0104	
206994_at	CST4		-0.0037		
207850_at	CXCL3		-0.0246		
208025_s_at	HMGA2	-0.0021		-0.0010	
209343_at	EFHD1				0.0096
212328_at	LIMCH1			0.0050	
213703_at	LINC00342	0.0008			
215867_x_at	CA12		-0.0026		
218677_at	S100A14				-0.0257
218824_at	PNMAL1	0.0003			
219654_at	PTPLA		-0.0240		
219747_at	NDNF	0.0002			
220952_s_at	PLEKHA5		-0.0047		
221841_s_at	KLF4	-0.0047			
222043_at	CLU		0.0049		

Table S2.4: Analysis of lung cancer data using GMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	DFCI	MSKCC
202503_s_at	KIAA0101	-0.0009	-0.0020	-0.0021	-0.0019
205776_at	FMO5	0.0001	0.0002	0.0002	-0.0001
207850_at	CXCL3	-0.0017	-0.0139	0.0029	0.0095
208025_s_at	HMGA2	-3.2E-05	1.1E-05	-3.8E-05	-2.2E-05
219654_at	PTPLA	-0.0036	-0.0092	-0.0024	0.0060
219764_at	FZD10	-0.0014	-0.0036	-0.0014	-0.0036